LigHTS: Lagging Hardware Tolerant Systems

THE UNIVERSITY OF CHICAGO

Haryadi Gunawi and Andrew Chien

UCARE ucare.cs.uchicago.edu

CERES ceres.uchicago.edu

(1) Limpware

Limpware

Definition: Hardware whose performance degrades significantly compared to its specs (a lagging hardware).

Anecdotal Impacts:
"… ***1Gb NIC*** *card on a machine that suddenly starts transmitting at* ***1 Kbps****, making the performance of entire workload for a* ***100 node cluster was crawling at a snail's pace***" – Facebook engineers.

A destructive failure mode:

- Cascading failures (entire cluster collapse)
- No "fail in place" recovery

Our findings

Current cloud systems are *susceptible to limpware.*

A single piece of limpware can cause *severe impact on a whole cluster* [SoCC '13].

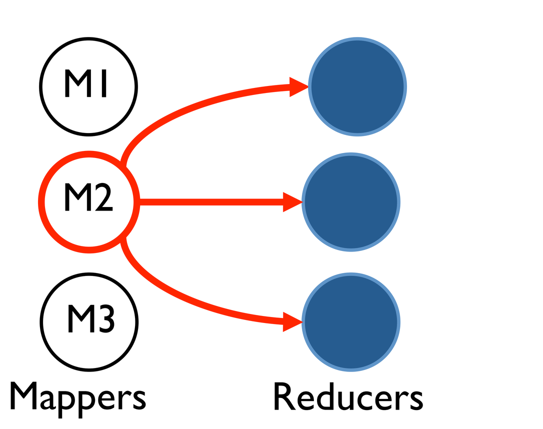
System	Operation	Node	Cluster
Hadoop	✖	✖	✖
HDFS	✖	✖	✖
ZooKeeper	✖	✖	✖
Cassandra			
HBase	✖	✖	

… even Hadoop speculative execution is not triggered!

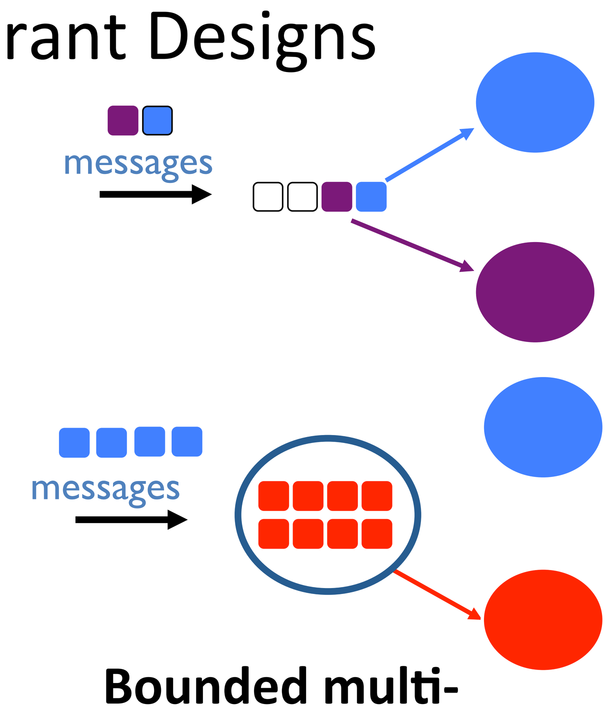
Limpware-Intolerant Designs

Problems:

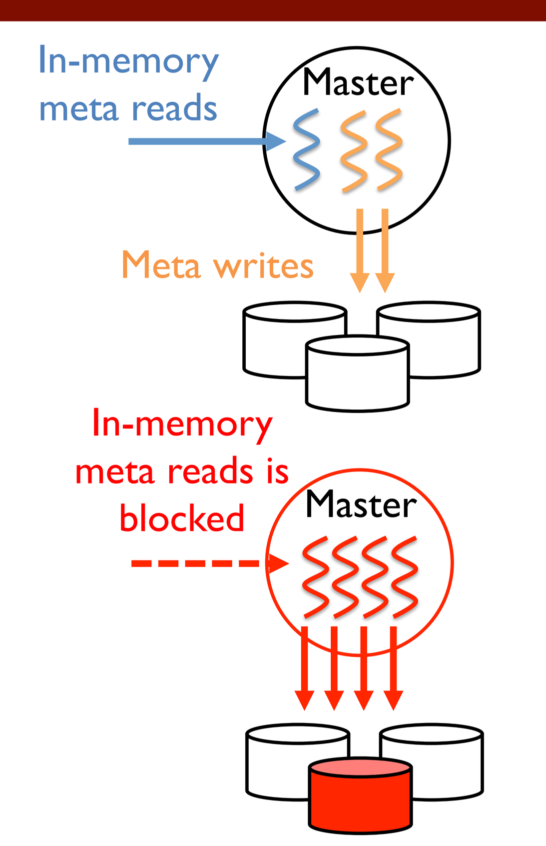
Limpware-Intolerant Designs



Single point of performance failure



Bounded multi-purpose queue



Bounded multi-purpose thread pool

Current work:

Pre-deployment Detection of Performance Bugs [HotCloud '15]

- Challenge: Various deployment scenarios such as data locality, data source, job characteristic, job/load size, fault type/placement/granularity/timing, topology scenario, etc.
- Solution: Convert (automatically) complex system code to formal model (CPN). Model check various deployment scenarios

Path-Informed Recovery

- Challenge: Limpware is not like fail-stop. Protocol callpath is deep (touches many hardware). Today's recovery sometimes cannot pinpoint the limpware.
- Solution: Ensure multi-layer systems manage paths. Recovery should not take the same slow path.

(2) The Tail at Store

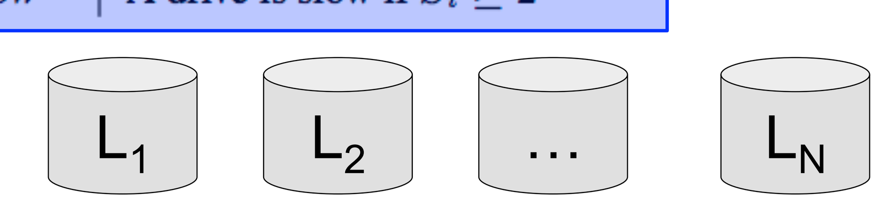
Large-scale Study of Storage Performance Variability

- Study of over **450,000 disks, 4000 SSDs, and 240 virtual drives** in deployment.
- More than **800 million drive hours** analyzed.
- (Collaboration with Gokul Soundararajan and Deepak Kenchammana of NetApp)

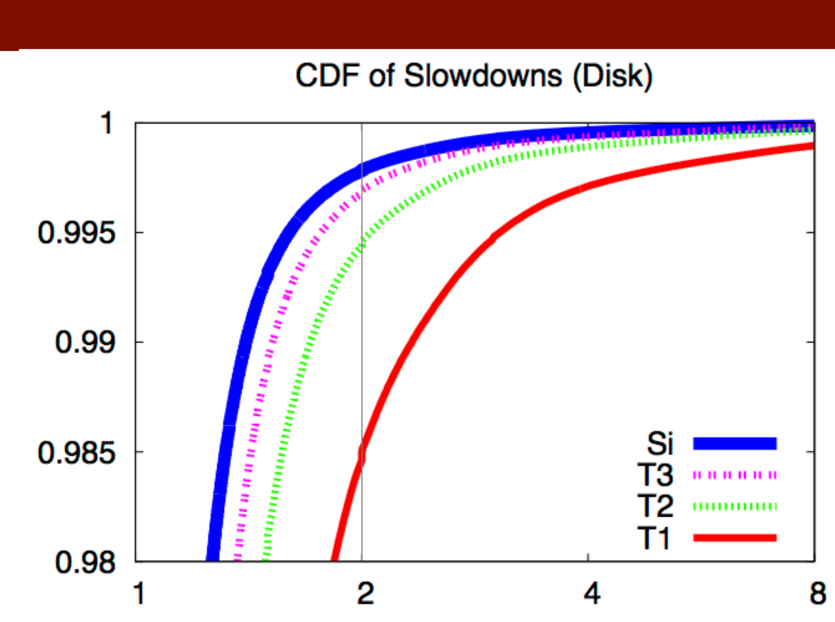
	Disk	SSD	Virtual
RAID groups	38,029	572	42
Data drives per group	3-26	3-22	3-23
Data drives	458,482	4,069	242
Slow drives (§3.3.4)	118,149	1,195	229
Duration (days)	1-1470	1-94	1-230
Drive hours	857,183,442	7,481,055	211,032
Slow drive hours (§3.3.1)	1,885,804	43,016	37,327
Slow drive hours (%)	0.22%	0.58%	17.7%
RAID hours	72,046,373	1,072,690	56,080
Slow RAID hours (§3.3.2)	1,109,514	23,964	31,230
Slow RAID hours (%)	1.54%	2.23%	55.7%

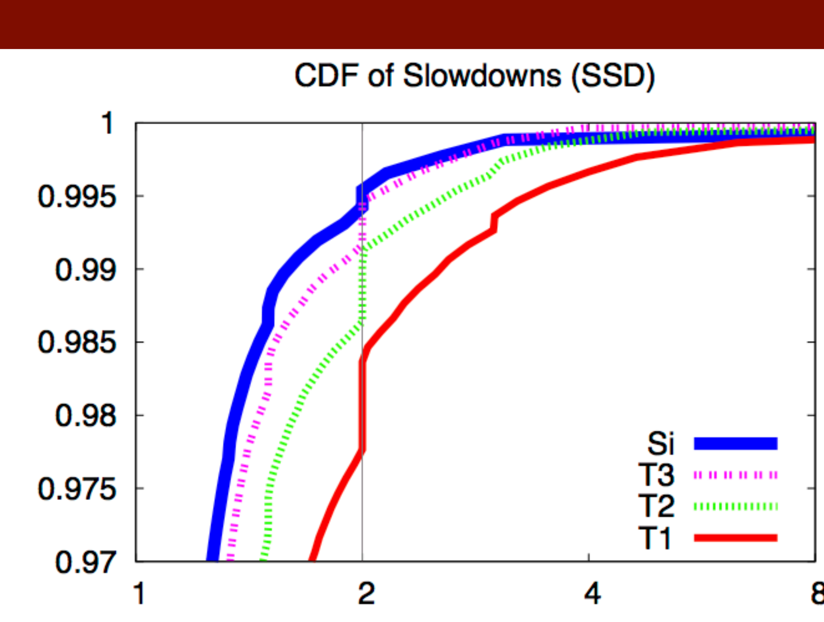
© 2015 NetApp, Inc. All rights reserved.

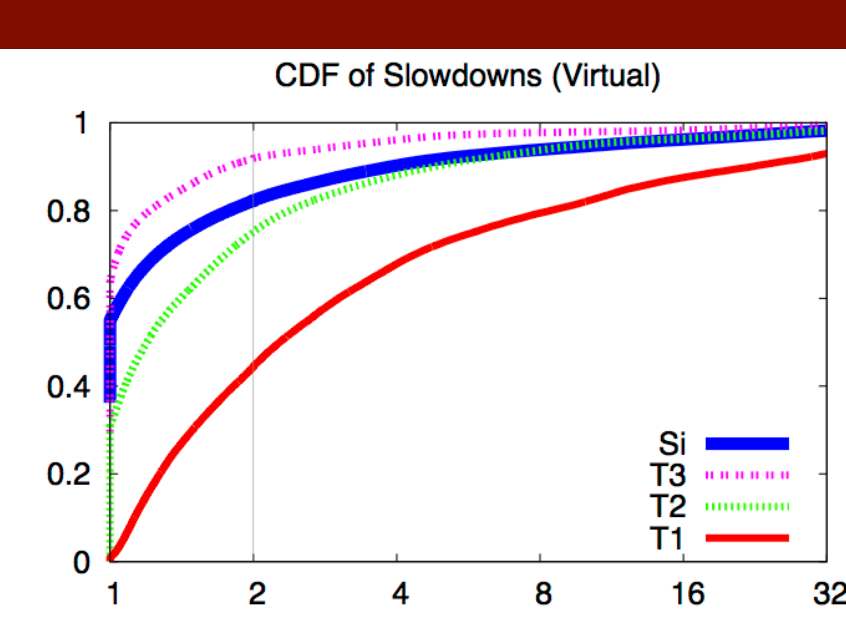
Label	Definition
N	Number of *data* drives in a RAID group
D_i	Drive number within a RAID group; $i = 1..N$
L_i	Hourly average I/O latency observed at D_i
L_{med}	Median latency; $L_{med} = Median\ of(L_{1..N})$
S_i	Latency slowdown of D_i compared to the median; $S_i = L_i/L_{med}$
T^k	The k-th largest slowdown ("k-th longest tail"); $T^1 = Max\ of\ (S_{1..N})$, $T^2 = 2nd\ Max\ of\ (S_{1..N})$, and so on
Stable	A drive is stable if $S_i < 2$
Slow	A drive is slow if $S_i \geq 2$

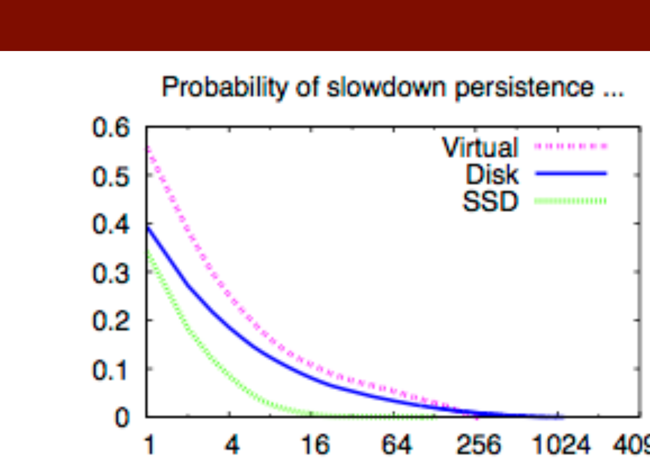


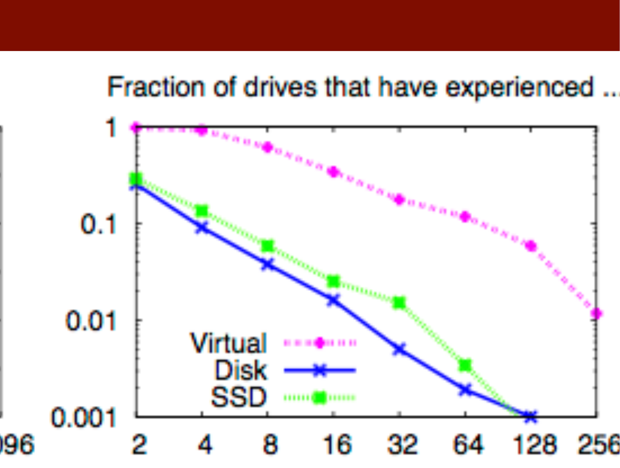
Results

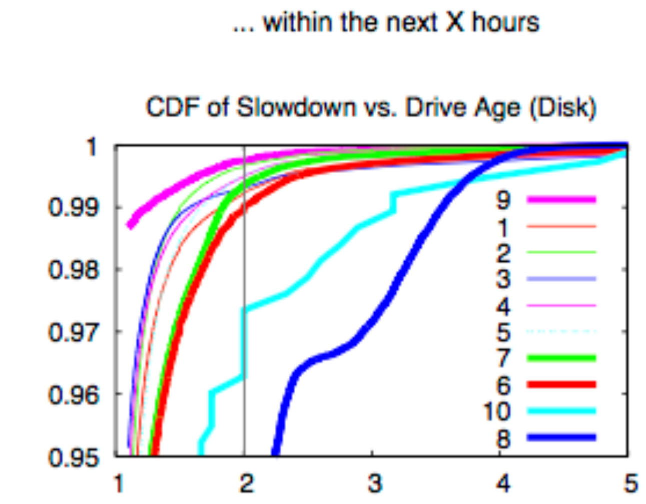


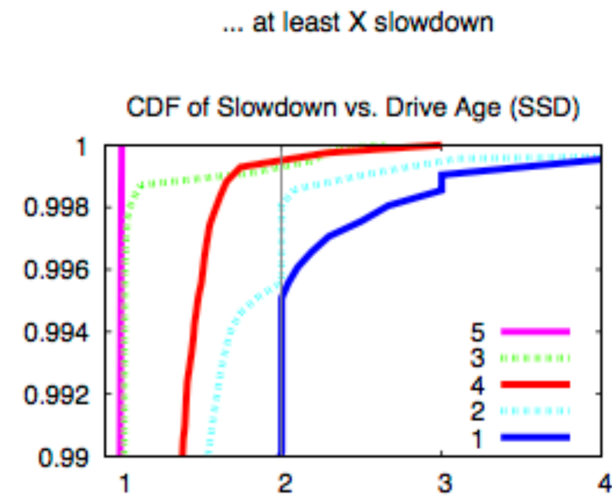












- 0.22% disk, 0.58% SSD and 17.7% vdrive hours experience > 2x slowdown
- 26% disks, 29% SSDs and 98% vdrives have experienced at least one slow hour in their lifetimes
- 40%, 35% and 55% of slow disks, SSDs, and virtual drives will stay slow within the next hour
- 2-digit slowdowns had occurred in 0.01% of disk and SSD hours, 4- and 3-digit slowdowns in 124 and 2461 disk hours, and 3-digit SSD slowdowns in 10 SSD hours